Statistica spaziale: dati e prime analisi

Carlo Gaetan

Dipartimento di Statistica
Università Ca' Foscari Venezia

Venezia, 2 dicembre 2009

Sommario

Introduzione: tipi di dati spaziali

Dati puntuali

Dati geostatistici

Dati di area

Esempi di dati spaziali (*dataset* in R)

1. Altezze piezometriche in una regione detta Wolfcamp Aquifer (Texas, USA)
2. Biomassa di *blue grama* rilevata in una zona di 200 x 200 metri vicino a Elgin (Arizona, USA)
3. Numero di morti dovuti a SIDS nelle contee dello stato del North Carolina (USA)
4. Valori dei pixel di una immagine satellitare
5. Posizione e diametri di alberi in una regione di 200 x 200 metri nel southern Georgia (USA)
6. Altri ...

Ingredienti di base I

- Le osservazioni provengono da ben identificati siti in una parte di uno spazio. Consideremo lo spazio geografico.
- La localizzazione di questi siti sono note e etichettano le osservazioni.
- Le osservazioni e/o le localizzazioni sono modellate come variabili casuali.
- Variabili risposta:
  - univariate, multivariate
  - continue, categoriali
  - a valori reali oppure no

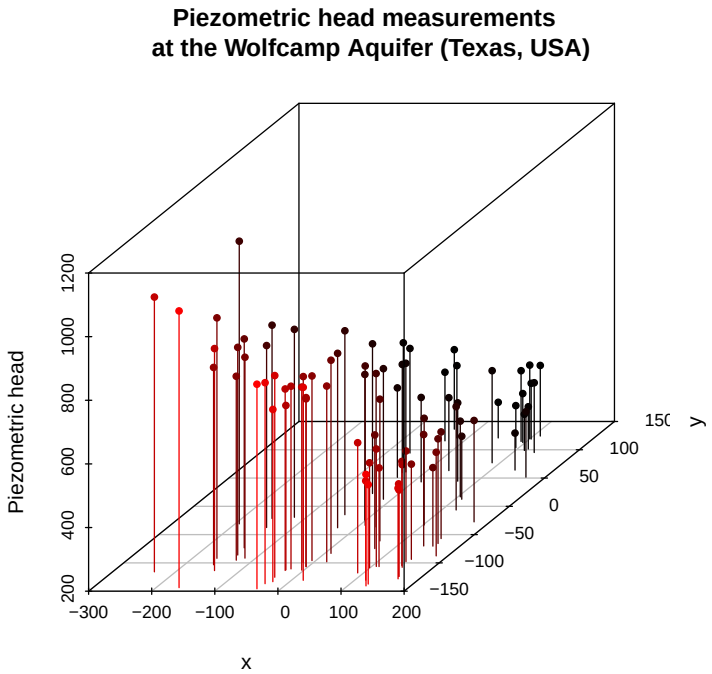
  localizzazione delle osservazioni sulla variabile risposta:
  - punti: regioni, segmenti, curve
  - irregolare, su una griglia
  - di forma regolare oppure no

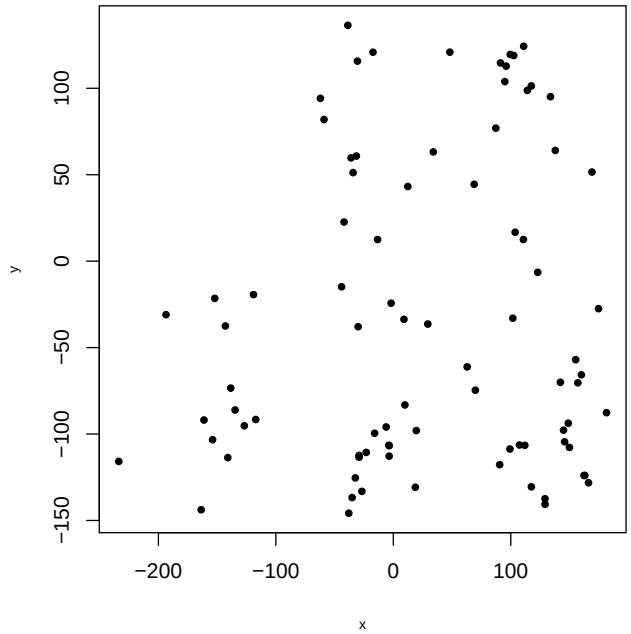
Ingredienti di base II

  - spazio euclideo oppure no
- Meccanismo di generazione della localizzazione
  - noto, non noto
  - casuale, non casuale
- Notazioni
  - spazio delle posizioni possibili: S (consideremo S come un piano)
  - punto arbitrario in S: s
  - regione di studio: D sottoinsieme di S
  - localizzazione di n osservazioni: $s_1, \ldots, s_n$
  - osservazioni: $Z(s_1), \ldots, Z(s_n)$

Dati geostatistici: altezze piezometriche

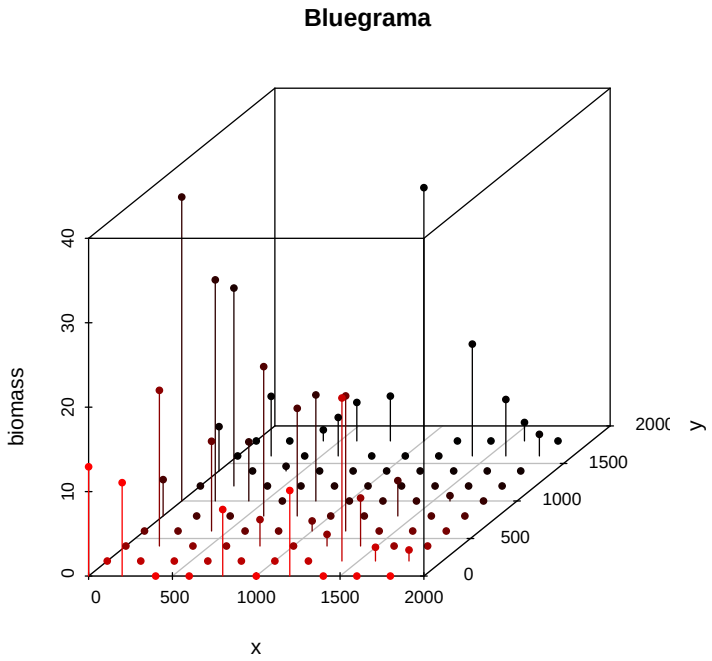
La variabile d'interesse esiste in ogni punto della regione ma si osserva solo la risposta in un insieme finito di localizzazioni.

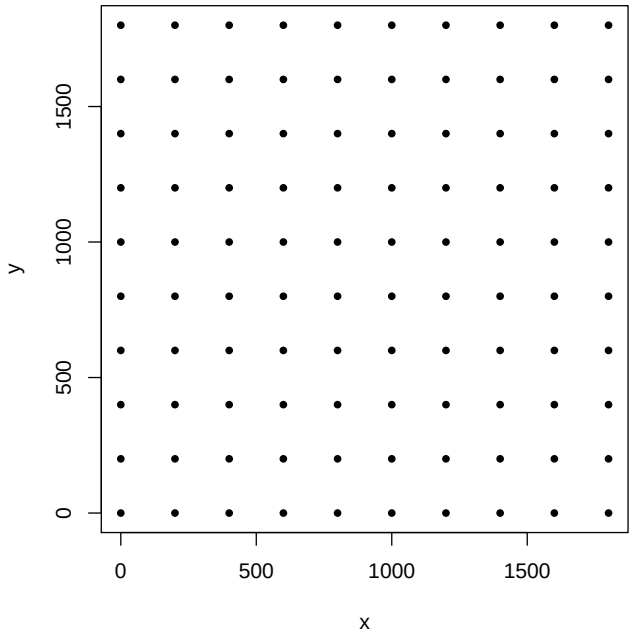




Dati geostatistici: biomassa del *blue grama*

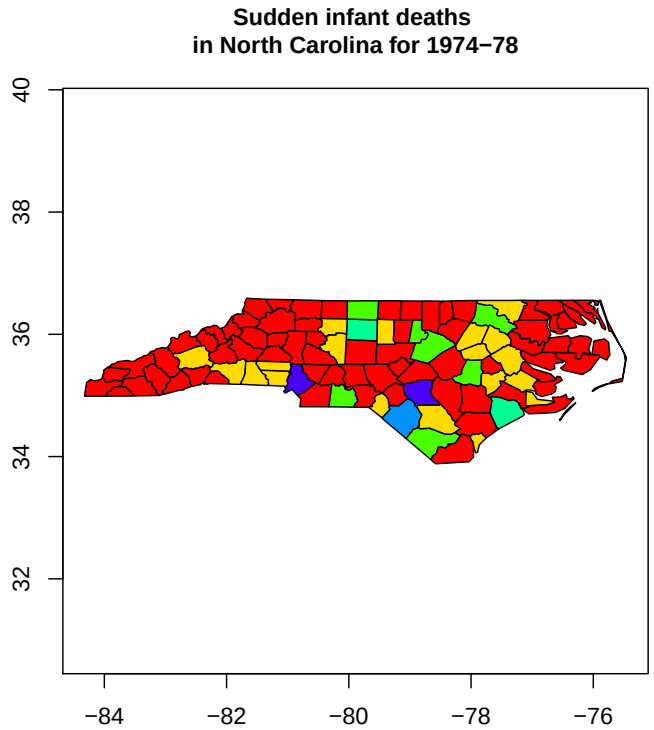
Si noti come la localizzazione possa essere regolare.

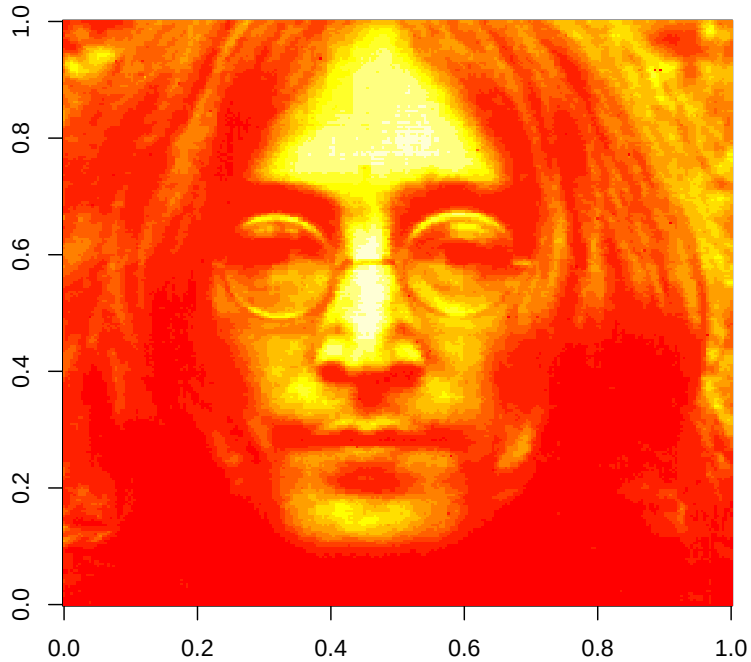




Dati su reticolo

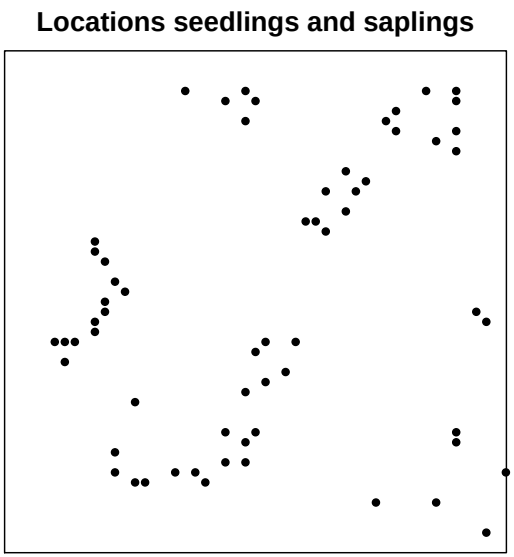
La variabile d'interesse esiste ed è osservata solo in un insieme finito di localizzazioni.

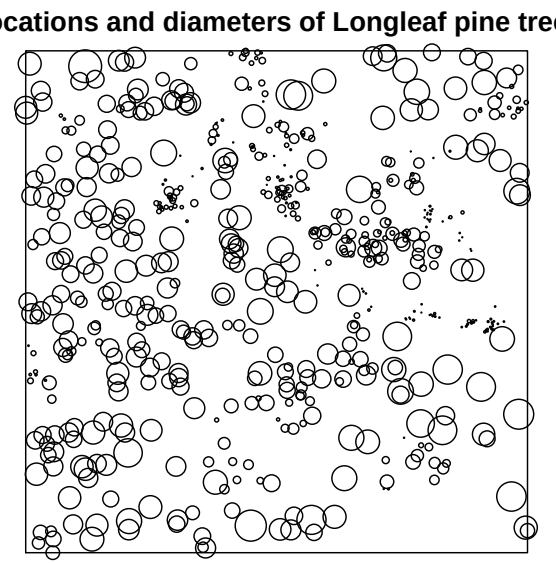




Dati puntuali

I dati sono le localizzazioni (supposte casuali). Qualche volta nelle localizzazioni si osserva una variabile risposta (*mark*), in questo caso il diametro.





Tipi di strutture spaziali I

Legge di Tobler

"*Osservazioni prese da siti vicini tendono ad essere più simili di osservazioni prese a siti distanti.*"

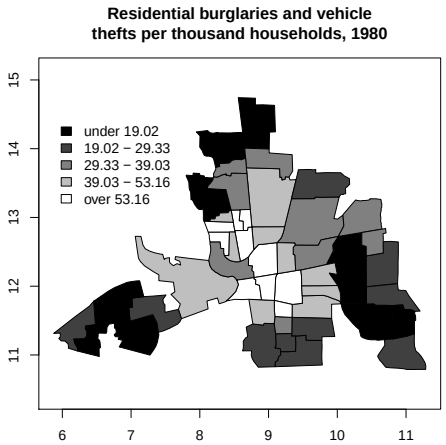
- Strutture di larga scala
  - Funzione media per dati geostatistici
  - Vettore delle medie per dati areali
  - Intensità per un processo di punto
- Strutture di piccola scala
  - Variogramma, funzione di covarianza
  - Matrice di vicinanza per dati areali
  - Funzione K di Ripley

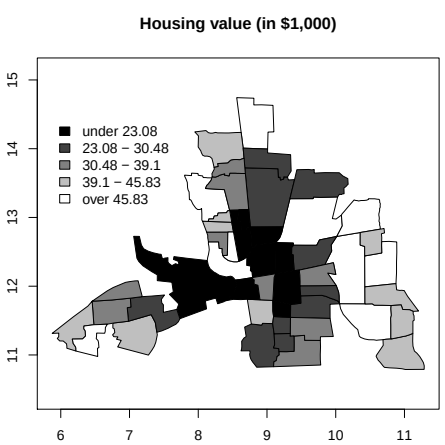
Tipi di strutture spaziali II

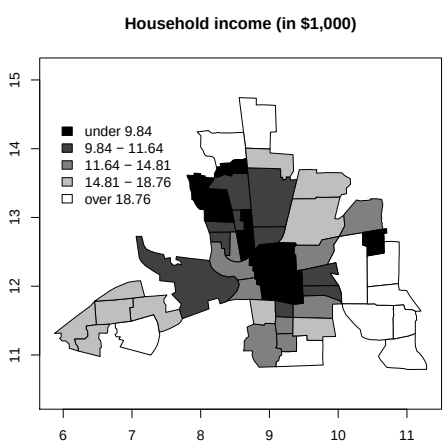
- Stazionarietà (il comportamento del processo generatore dei dati è simile in tutti i sottoinsiemi di S)
  - struttura di larga scala costante
  - struttura di piccola scala che dipende dalla posizione relativa
- Isotropia (il processo è stazionario e la struttura di piccola scala dipende solo dalla distanza tra siti).

Alcuni obiettivi della statistica spaziale I

- Inferenza sulla struttura spaziale (verifica sull'esistenza di una struttura spaziale, stima di questa). Importante per la validità delle procedure statistiche !
- Inferenza sulla struttura non spaziale (stima degli effetti di un trattamento, effetti delle covariate, stima del numero di punti)





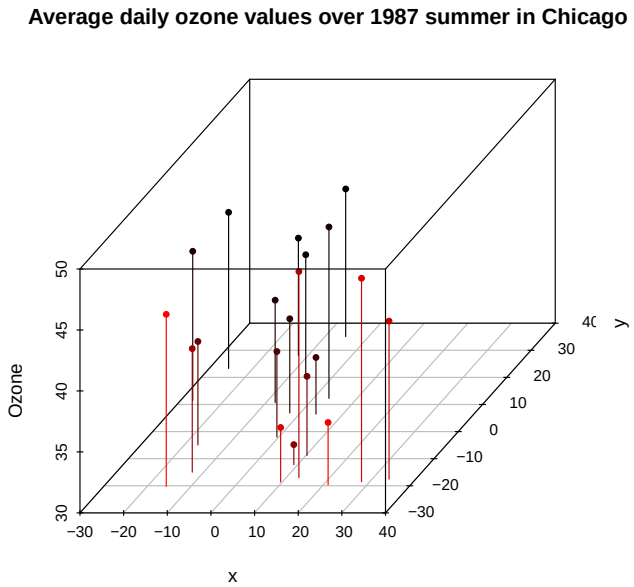


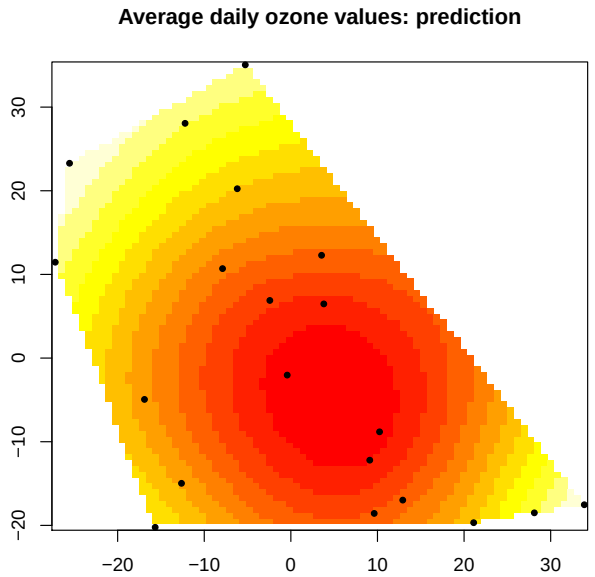
Z_i $HOVAL_i$ INC_i

$$Z_i = \beta_0 + \beta_1 INC_i + \beta_2 HOVAL_i + \eta_i$$

Alcuni obiettivi della statistica spaziale II

- Previsione di variabili non osservate (*kriging*, disegno spaziale ovvero come scegliere nuovi siti per nuove osservazioni ovvero come riposizionare i siti esistenti)

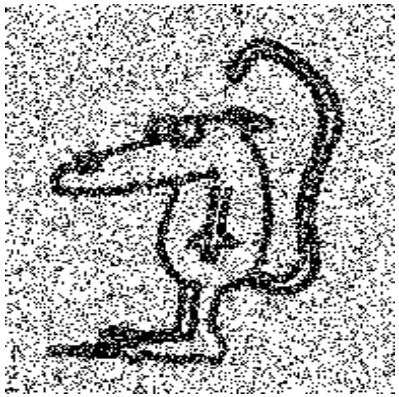




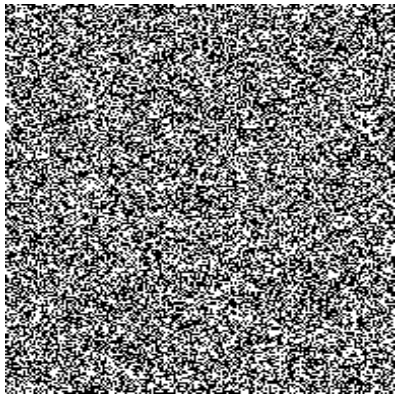
Alcuni obiettivi della statistica spaziale III

- estrazione di segnale, ricostruzione di immagini

x



z



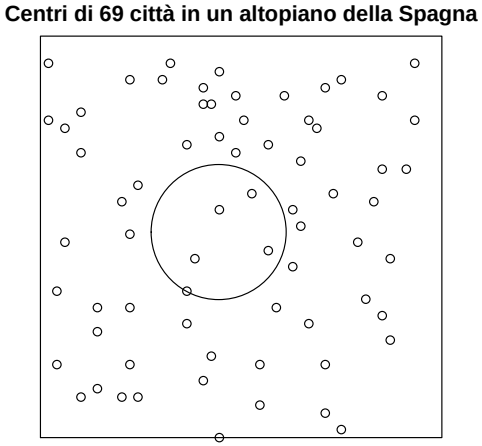
$\hat{z}(0)$

$\hat{z}$

Dati puntuali

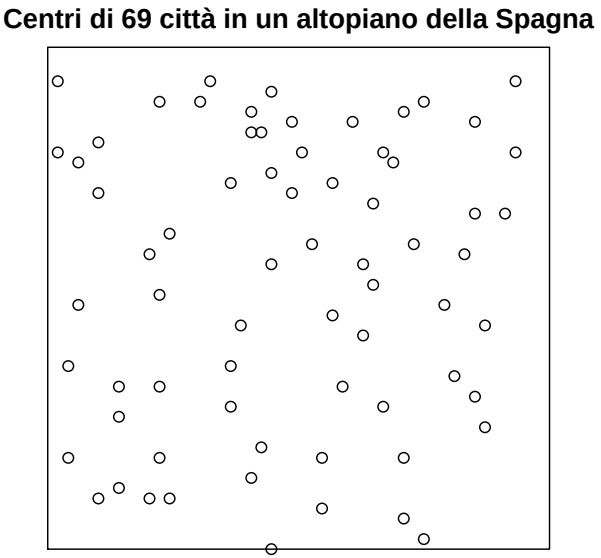
Processi di punto

- Un processo di punto è un meccanismo generatore che determina la posizione di un insieme di n punti $s_1, \ldots, s_n$ nello spazio.
- Da un punto di vista probabilistico un processo di punto può essere caratterizzato dal numero di punti che cadono in una regione prefissata A, $N(A)$. Questo numero è una variabile casuale. L'area di A sarà indicata con $|A|$.



Esempi di configurazioni spaziali I

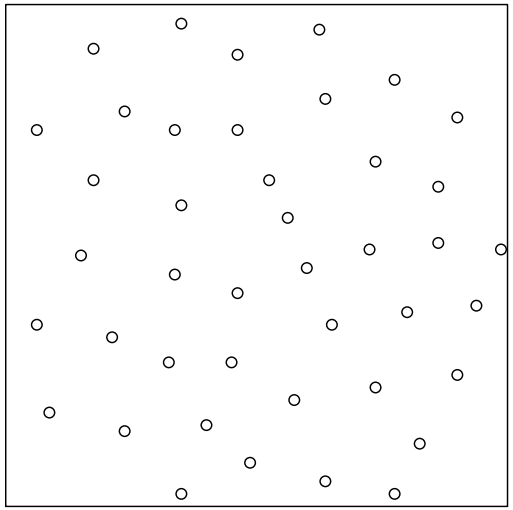
Casualità completa



Esempi di configurazioni spaziali II

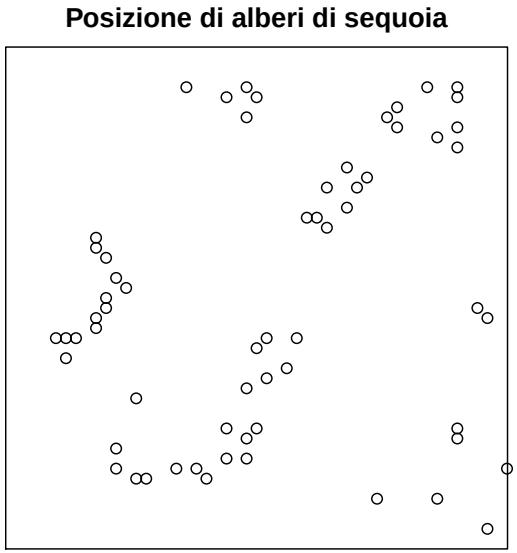
Regolarità

Cellule di una sezione istologica



Esempi di configurazioni spaziali III

Aggregazione



Caratteristiche di un processo di punto I

- $N(A)$ è una variabile casuale con una certa media $\mathrm{E}(N(A))$
- Questa media varia con A. Se consideriamo una regione di area infinitesima ds possiamo definire l'intensità del processo come

$$\lambda(s) = \lim_{|ds| \to 0} \frac{E(N(ds))}{|ds|}$$

  Tale funzione è detta **funzione d'intensità del primo ordine**.
- Un processo **stazionario** è tale per cui $\lambda(s)$ assume un valore costante λ su tutta la regione ; quindi $E(N(A)) = \lambda|A|$

Caratteristiche di un processo di punto II

- La **funzione di intensità di secondo ordine** caratterizza la dipendenza spaziale, ed è definita come

$$\lambda_2(s_i, s_j) = \lim_{|ds_i|,|ds_j|\to 0} \frac{E(N(ds_i)E(N(ds_j)}{|ds_i||ds_j|}$$

- Per un processo stazionario,

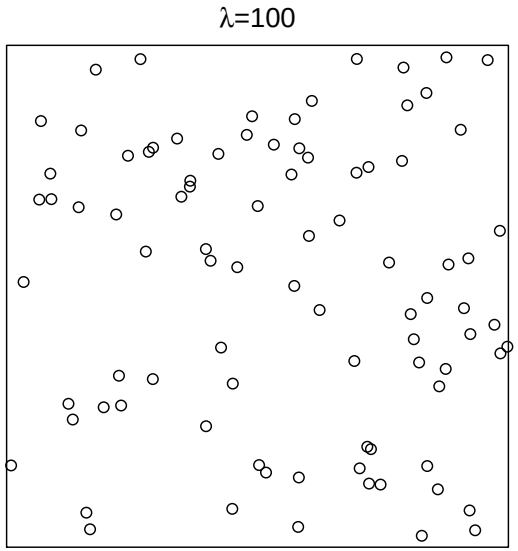
$$\lambda_2(s_i, s_j) = \lambda_2(s_i - s_j) = \lambda_2(h)$$

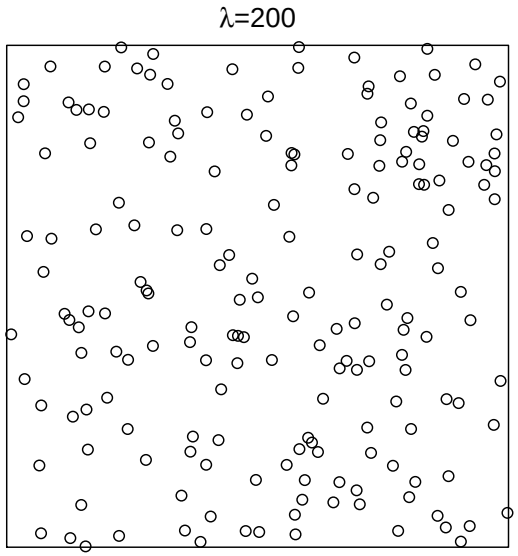
con h il vettore differenza tra s_i ed s_j per un processo stazionario.

Casualità spaziale completa e Processo di Poisson omogeneo I

- Si consideri una regione prefissata A.
- La casualità spaziale completa (*Complete Spatial Randomness*) nella regione A corrisponde alla generazione di un processo di Poisson omogeneo.
- Un processo di Poisson omogeneo d'intensità λ ha due caratteristiche:
  1. per ogni n punti nella regione "i punti si distribuiscono uniformemente in A"
  2. il numero di punti nella regione A ha una distribuzione di Poisson con media $\lambda|A|$ ($|A|$ è l'area di A)

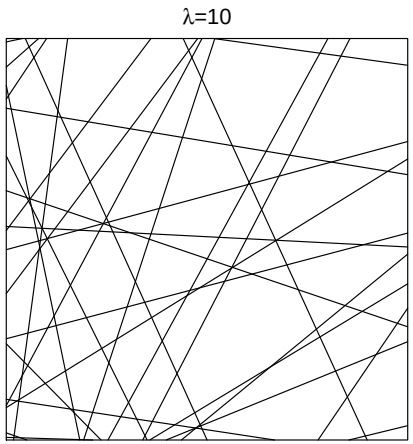
Casualità spaziale completa e Processo di Poisson omogeneo II



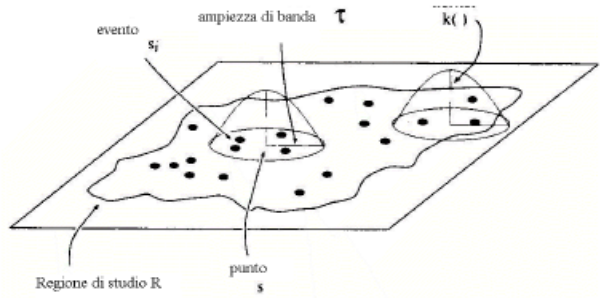


Casualità spaziale completa e Processo di Poisson omogeneo III

Il numero atteso di linee che si intersecano una regione convessa del piano A è pari a $\lambda p(A)$, $p(A)$ è il perimetro di A. La lunghezza attesa totale delle linee che attraversano una regione del piano è pari a $\lambda\pi|A|$



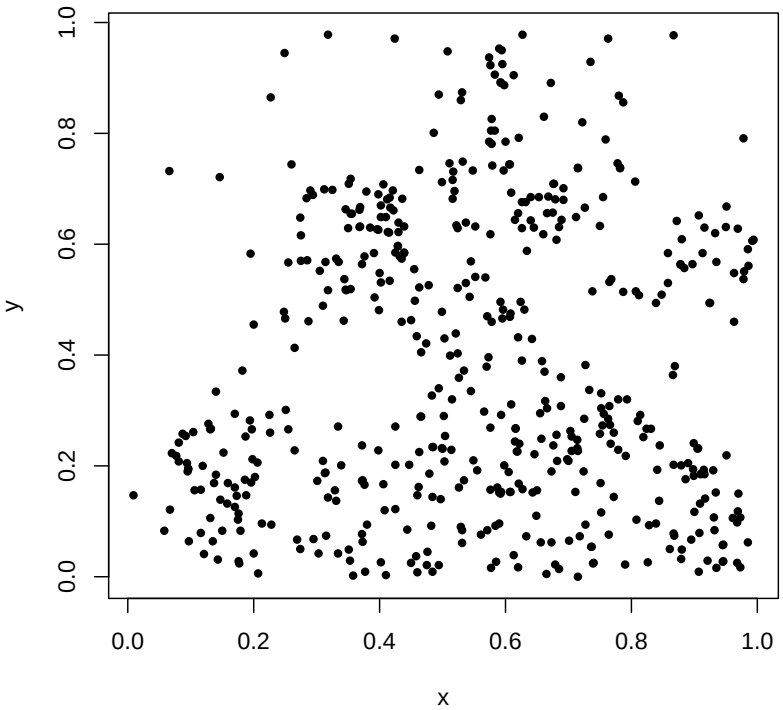
Stima non parametrica della funzione di intensità I

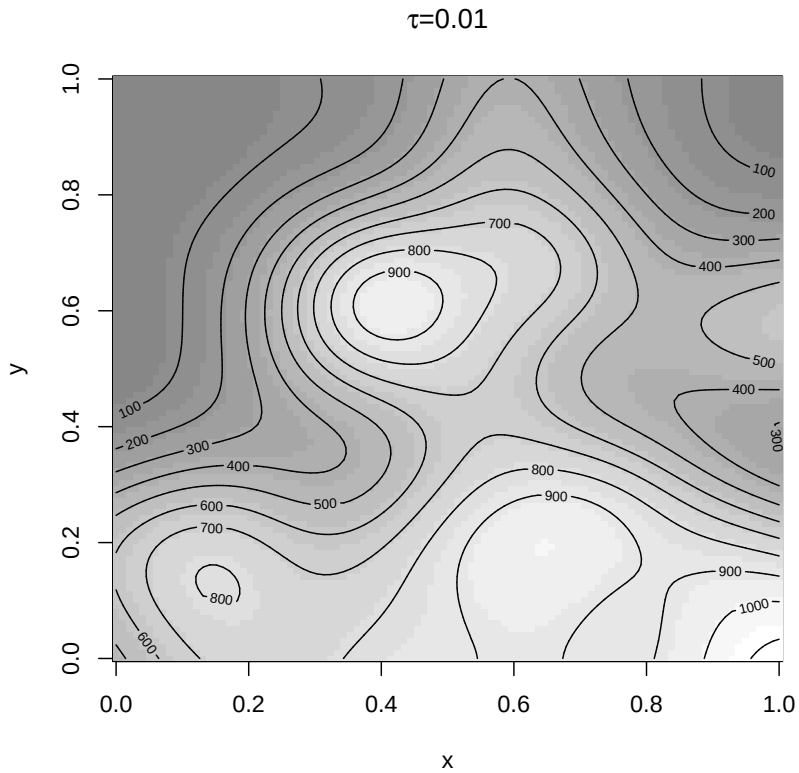


$$\hat{\lambda}(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} K\left(\frac{s - s_i}{\tau}\right)$$

dove $K(\cdot)$ è una densità di probabilità bivariata simmetrica rispetto l'origine, detta *kernel*. Il parametro $\tau > 0$ è noto come ampiezza di banda (*bandwidth*) e determina il grado di "lisciamento" dell'intensità stimata (rappresenta il raggio del disco centrato su s entro cui i punti s_i contribuiscono alla stima di $\lambda(s)$).

Stima non parametrica della funzione di intensità II





Posizione di aceri in un terreno di 19.6 acri (Lansing Woods, Clinton County, Michigan USA)

Funzione K di Ripley I

- La funzione K di Ripley

$$K(h) = \frac{\mathrm{E}(N_0(h))}{\lambda}$$

dove

  - $N_0(h)$ è il numero degli eventi che stanno dentro un cerchio di raggio h con centro un evento arbitrario
  - λ è l'intensità del processo.

Funzione K di Ripley II

- Nel caso di un processo di Poisson omogeneo (CSR)

$$K(h) = \pi h^2$$

infatti in un cerchio di raggio h in media abbiamo un numero atteso di punti

$$\mathrm{E}(N_0(h)) = \lambda \times \text{'Area cerchio di raggio } h\text{'} = \lambda \pi h^2$$

- Nel caso di aggregazione $K(h) > \pi h^2$
- Nel caso di regolarità $K(h) < \pi h^2$

Funzione K di Ripley III

LA stima di K è data da

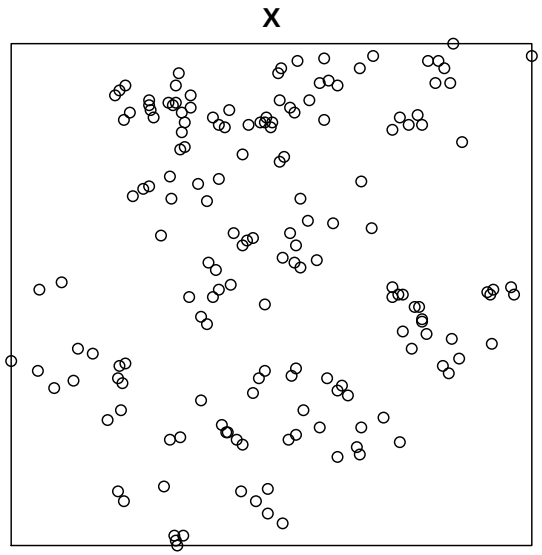
$$\widehat{K}(h) = \frac{1}{\lambda^2 |A|}\{\#\text{di coppie } (s_i, s_j) \text{ con distanza} \leq h\}$$

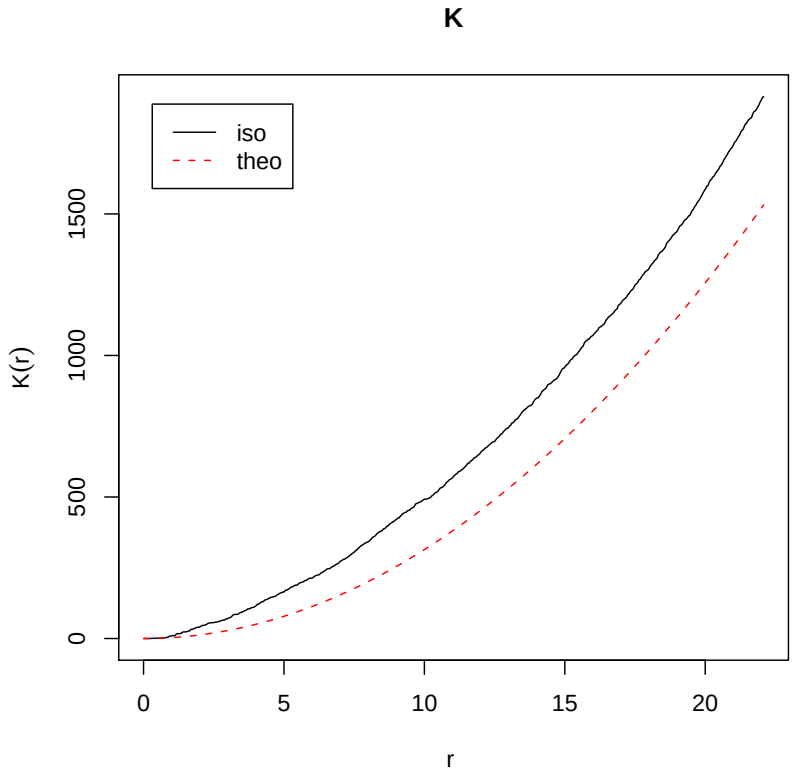
una stima per λ è

$$\widehat{\lambda} = \frac{n}{|A|}$$

Funzione K di Ripley IV

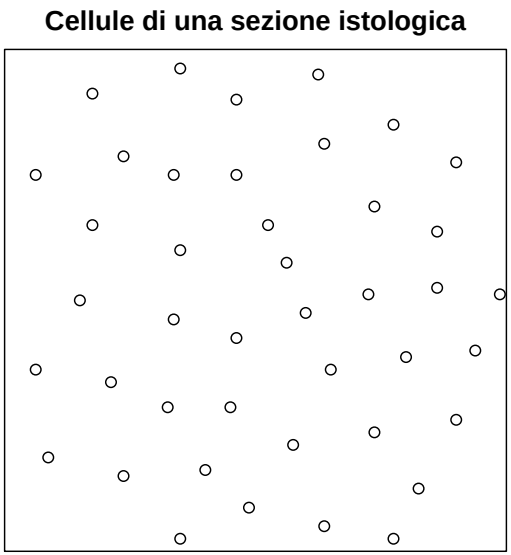
Residenze di giovani criminali, i cui crimini sono stati registrati nel 1971 a Cardiff (Galles, UK).

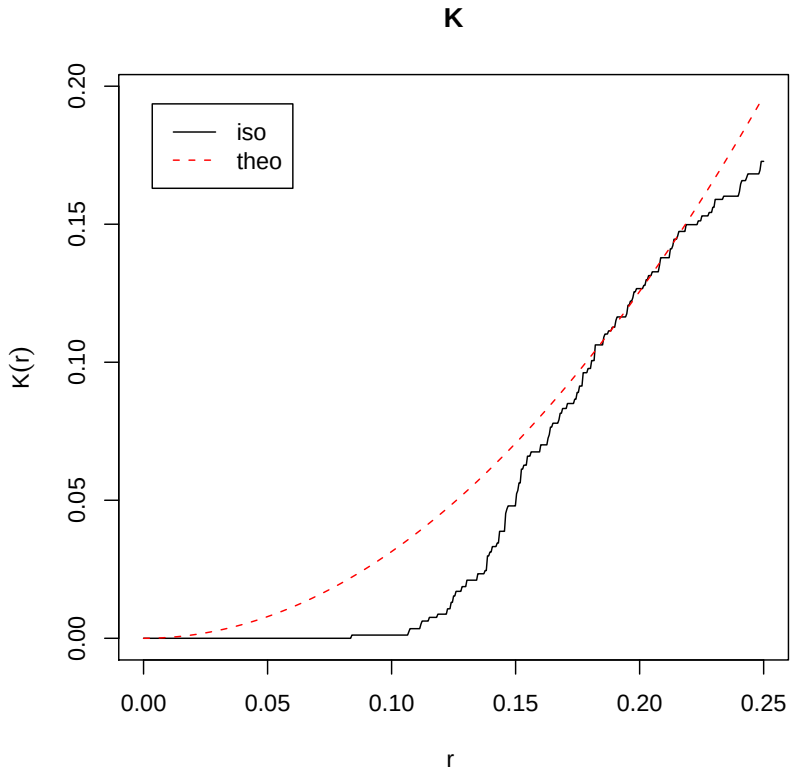




Presenza di aggregazione !

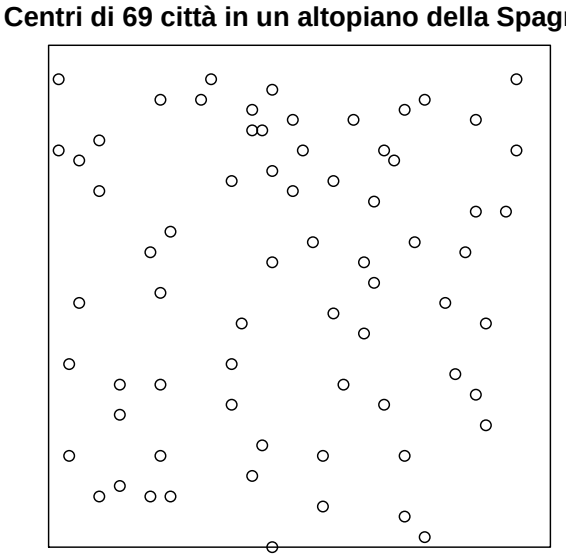
Funzione K di Ripley V

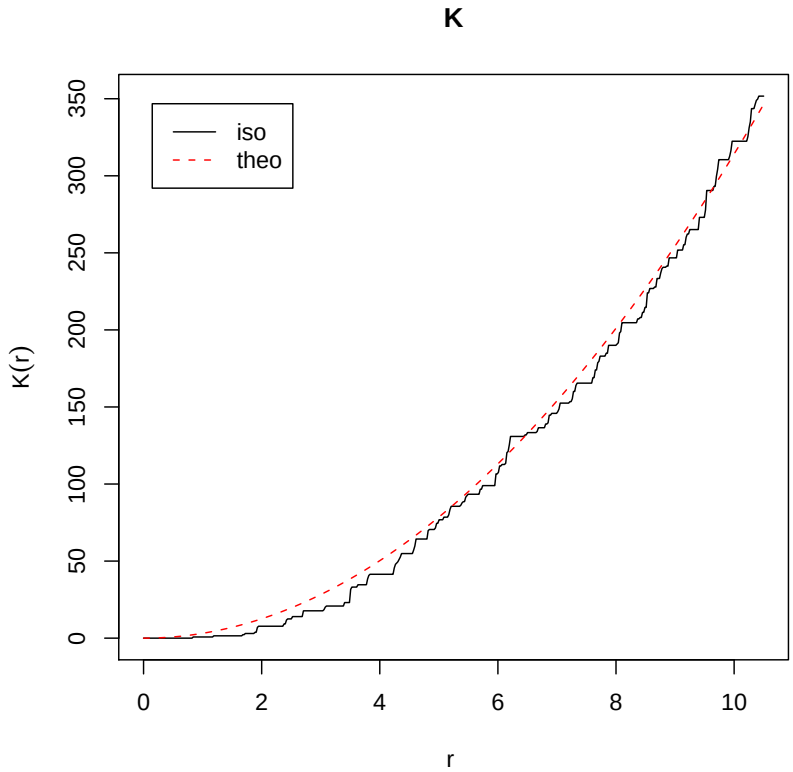




Presenza di regolarità

Funzione K di Ripley VI

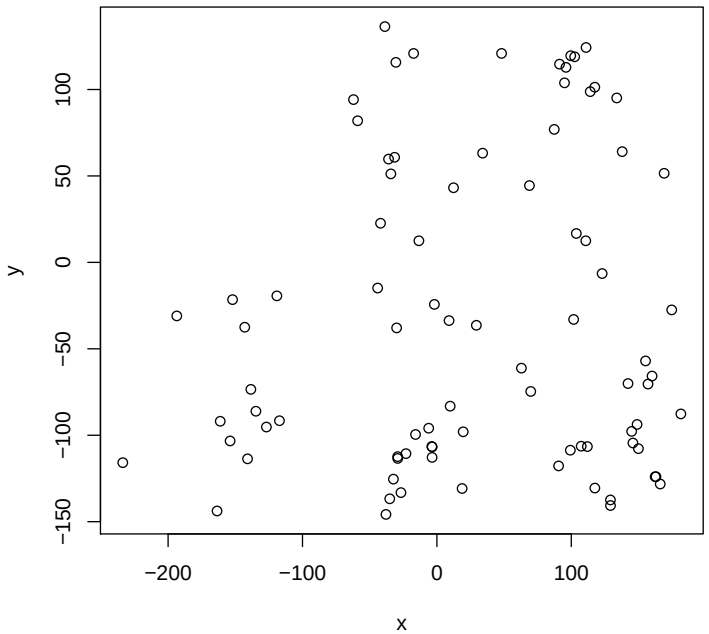




Dati geostatistici

Notazioni

Considereremo una regione S che è un sottoinsieme di $\mathbb{R}^2$ (per semplicità), un punto s avrà coordinate $s = (x, y)'$. In n punti distinti $s_1, \ldots, s_n$



vengono considerate n osservazioni $Z(s_1), \ldots, Z_s(n)$

Piezometric head measurements

La stima della componente di larga scala I

Si supponga che

$$Z(s) = \mu(s) + \varepsilon(s)$$

con $\varepsilon(s)$ componente d'errore. Si vuole trovare una stima $\widehat{\mu}(s)$ di $\mu(s)$, mediante le n osservazioni $\mathbf{Z} = (Z_(s_1), \ldots, Z(s_n))'$ disponibili. Per questo consideremo vari metodi.

Media aritmetica

- Ipotesi: Si suppone che la superficie $Z(\cdot)$ si muova attorno ad un valore 'centrale' (μ 'costante').
- Stima

$$\widehat{\mu}(s) = \sum_{i=1}^{n} Z(s_i)/n$$

o

$$\widehat{\mu}(s) = \mathrm{med}\{Z_(s_1), \ldots, Z(s_n)\}'$$

La stima della componente di larga scala II

- Proprietà: semplicità, velocità di calcolo.

Media mobile

- Ipotesi: si suppone che la superficie $\mu(\cdot)$ sia "sufficientemente liscia".
- Stima

$$\widehat{\mu}(s) = \sum_{i=1}^{n} Z(s_i) I(d_{s,s_i} \leq r) / \sum_{i=1}^{n} I(d_{s,s_i} \leq r)$$

dove $d_{s,s_i} = \|s_0 - s_i\|$, oppure ($k$ vicini) se $d_{(s,s_1)} \leq \cdots \leq d_{(s,s_n)}$

$$\widehat{\mu}(s) = \sum_{i=1}^{n} Z(s_i) I(d_{s,s_i} \leq d_{(s,s_k)}) / k$$

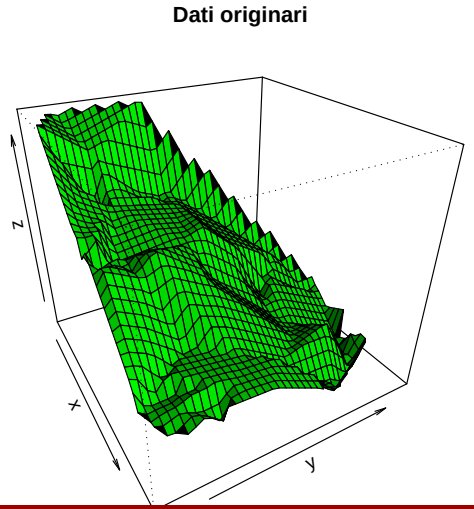
La stima della componente di larga scala III

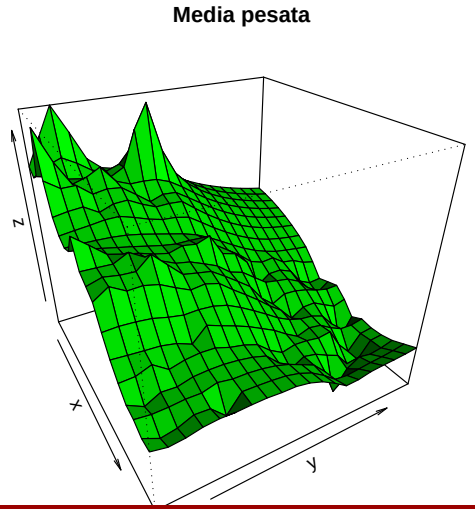
- Proprietà: è necessario scegliere r o k (grado di lisciamento), richiede algoritmi di ordinamento.

Medie pesate seconde le distanze Le singole osservazioni non contribuiscono più in maniera uguale.

$$\widehat{\mu}(s) = \sum_{i=1}^{n} d_{s,s_i}^{-p} Z(s_i) / \sum_{i=1}^{n} d_{s,s_i}^{-p}, \qquad p \geq 0$$

Proprietà: semplicità, velocità di calcolo





Covarianza e Covarianza campionaria

Date due variabili casuali (non spaziali) X e Y

- Covarianza:

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))]$$

$$\mathrm{Cov}(X, X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{Var}(X) \qquad \textbf{(varianza)}$$

- Covarianza campionaria: date n osservazioni $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_n\}$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

$s_{xx} = s_x^2$ **(varianza campionaria)**.

Correlazione e Correlazione campionaria

Date due variabili casuali (non spaziali) X e Y

- Correlazione:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  misura standardizzata $-1 \leq \mathrm{Corr}(X, Y) \leq 1$

- Correlazione campionaria:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

  misura standardizzata $-1 \leq r_{xy} \leq 1$

Auto covarianza

- Covarianza di una variabile con se stessa ?
- Covarianza di una variabile osservata in un sito (indichiamolo con s) con una variabile osservata in un altro (indichiamolo con s')

$$\mathrm{Cov}(Z(s), Z(s')) = \mathrm{E}\{[Z(s) - \mathrm{E}(Z(s))][Z(s') - \mathrm{E}(Z(s'))]\}$$

Auto covarianza campionaria

Consideriamo un campione $\{z(s_1), \ldots, z(s_n)\}$

- Covarianza di una variabile osservata in un sito (indichiamolo con s_i) con una variabile osservata in un altro (indichiamolo con s_j). Ammettiamo che abbiano la stessa media campionaria $\overline{z}$.

$$s_{s_i,s_j} = (z(s_i) - \overline{z})(z(s_j) - \overline{z})$$

- $n(n-1)/2$ coppie (tante!)
- non molto utile (ovvero molto variabile) in quanto si riferisce ad una sola coppia !
- potremmo pensare di raccogliere insieme le coppie con le stesse caratteristiche ad esempio le coppie i cui siti sono separati dalla medesima quantità $s_i - s_j = h$

$$s_h = \frac{\sum_{s_i - s_j = h}(z(s_i) - \overline{z})(z(s_j) - \overline{z})}{N(h)}$$

(Semi) varianza campionaria delle differenze

E' calcolata non sui dati ma su tutte le possibili differenze

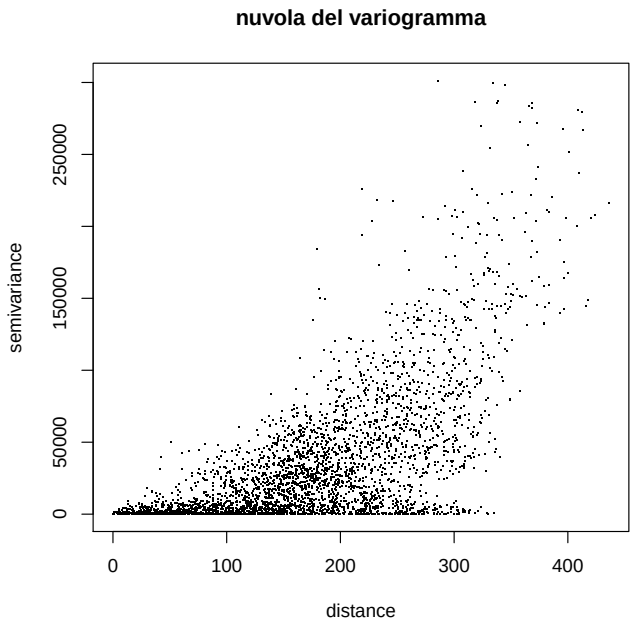
$$\frac{(z(s_i) - z(s_j))^2}{2}$$

- è una misura di dissimilarità
- coppie vicine nello spazio dovrebbero avere valori simili e quindi differenze vicine allo zero

Nuvola del variogramma I

Diagramma di dispersione delle coppie

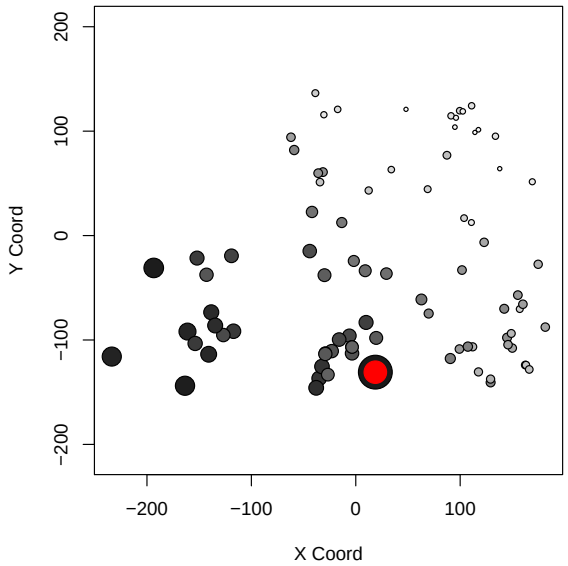
$$s_i - s_j, \frac{(z(s_i) - z(s_j))^2}{2}$$

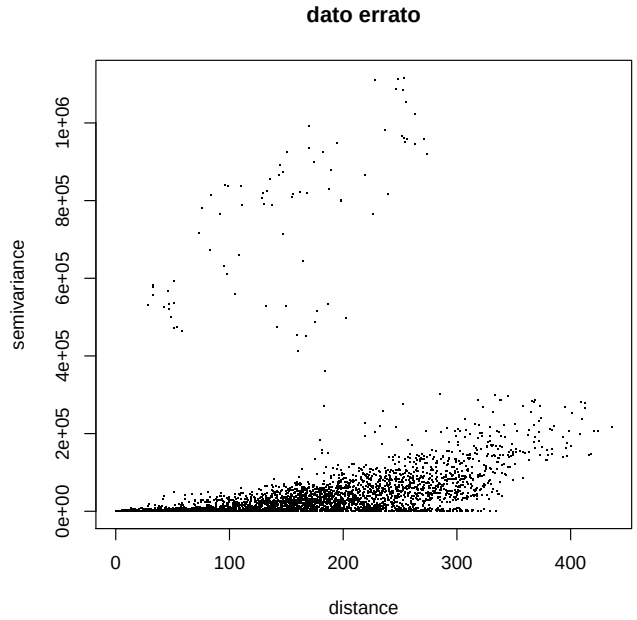


Nuvola del variogramma II

Utile nel rilevare

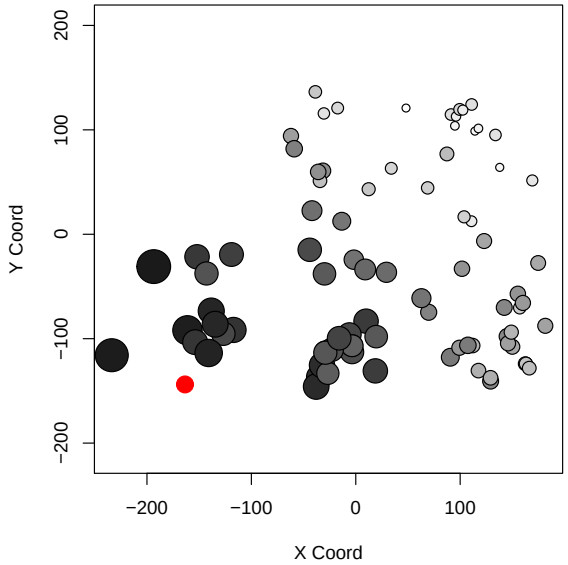
- valori anomali globali;

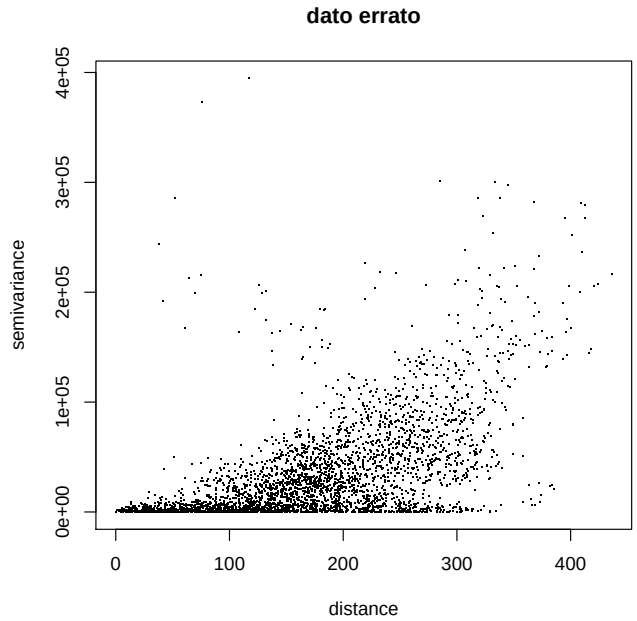




Nuvola del variogramma III

- valori anomali locali;



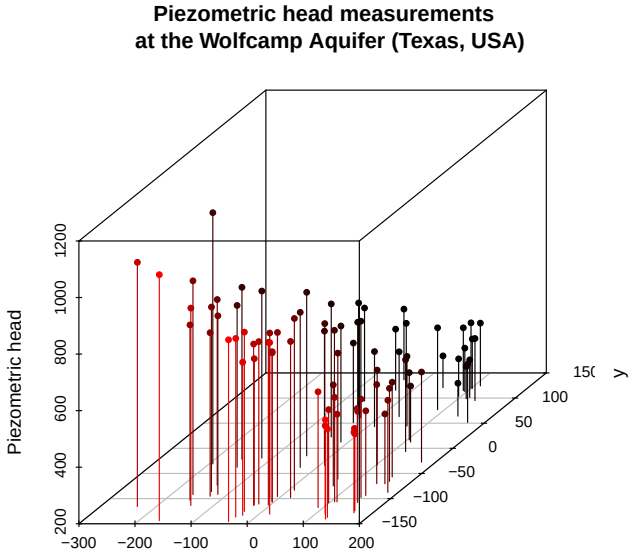


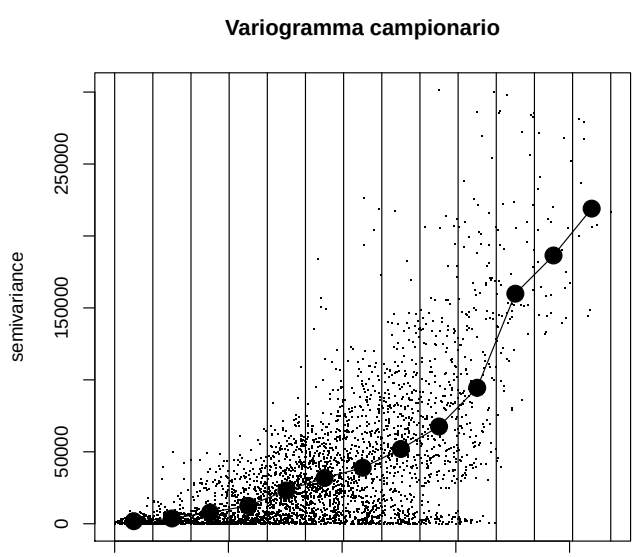
Variogramma campionario

Il semi-variogramma campionario riassume la nuvola del variogramma

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{s_i - s_j = h} (z(s_i) - z(s_j))^2$$

Nel caso in cui i siti non siano disposti su di un reticolo dobbiamo creare dei contenitori (*bins*) per contenere le coppie che sono separate approssimativamente da h

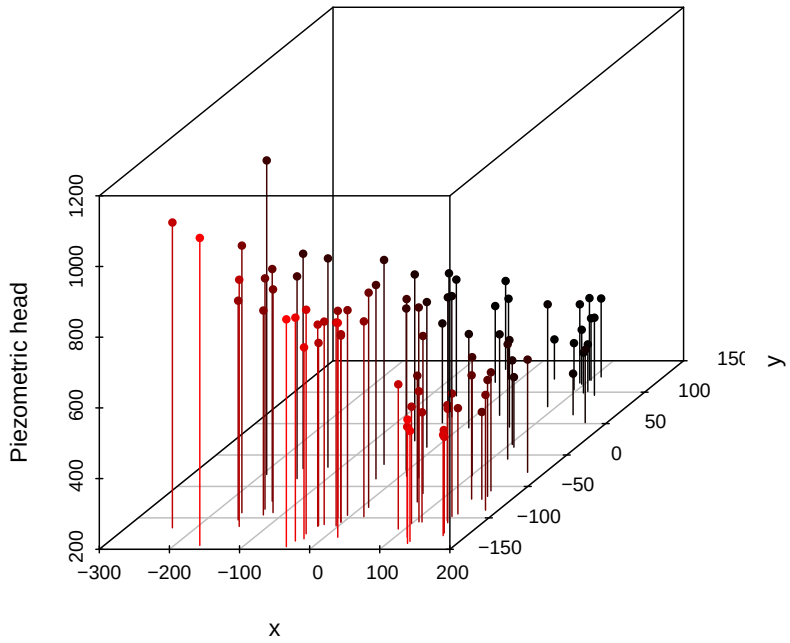




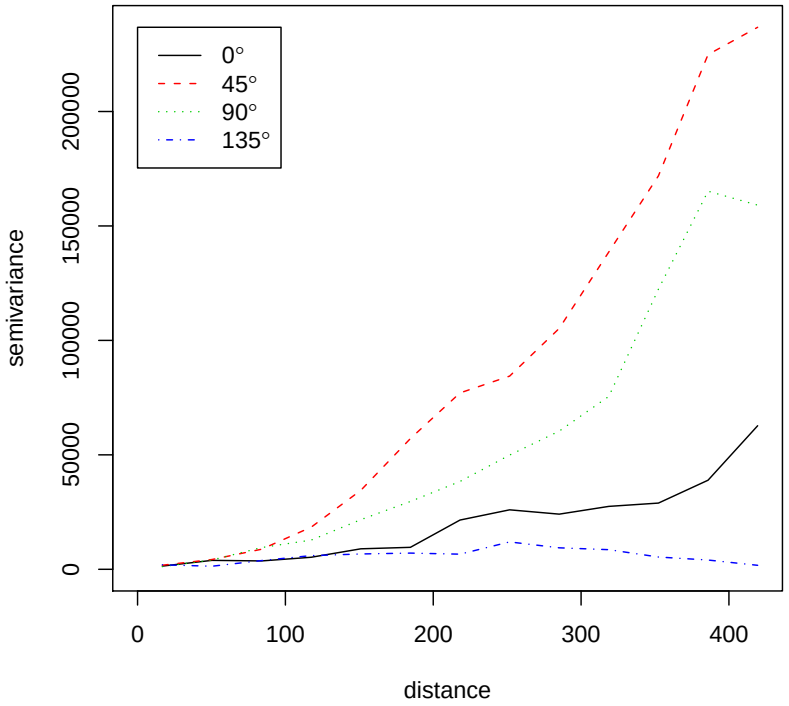
Il variogramma nelle varie direzioni

Anisotropia: la variabilità dipende dalla direzione

Piezometric head measurements at the Wolfcamp Aquifer (Texas, USA)

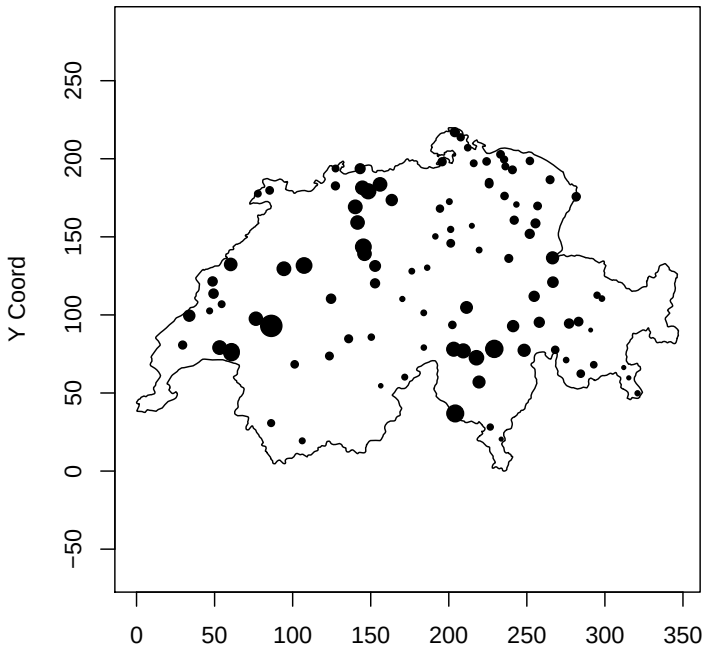


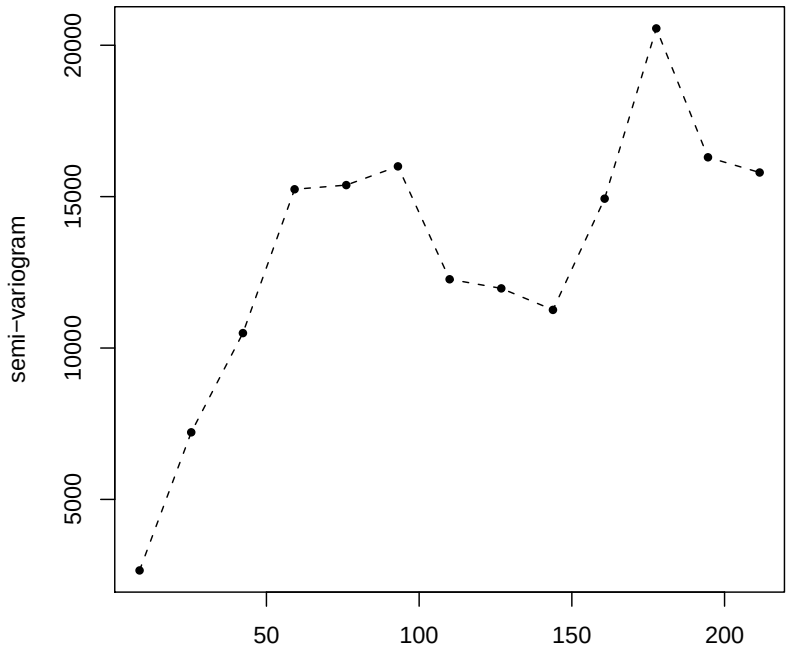
Variogramma campionario nelle quattro direzioni : N−S, NE−SW, E−W e SE−NW



Esempio: precipitazioni in Svizzera

I dati si riferiscono alla pioggia caduta sulla Svizzera l'8 maggio 1986, giorno in cui la nube di Chernobyl si è trovata al di sopra dell'Europa centrale. I dati sono stati oggetti di una competizione per trovare il miglior previsore di questi.





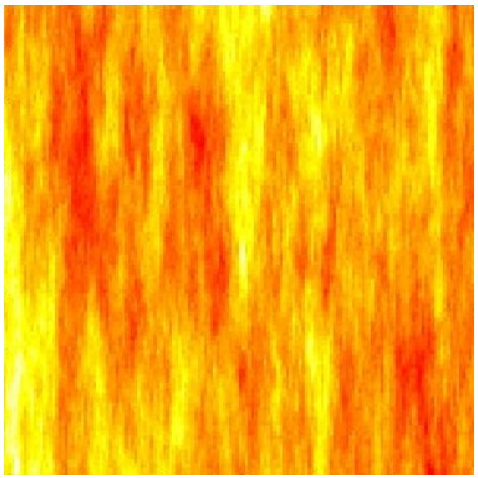
Modelli per la covarianza I

- Consideriamo la covarianza tra due siti $\mathrm{Cov}(Z(s), Z(s'))$ in generale questa dipenderà da s e s'.
- Supponiamo ora che $Z(s)$ e $Z(s')$ abbiano la stessa media e che la covarianza dipenda solo da $h = s - s'$

$$C(s - s') = \mathrm{Cov}(Z(s), Z(s')).$$

Tale ipotesi è detta **stazionarietà**,

Campo di colza in Lombardia

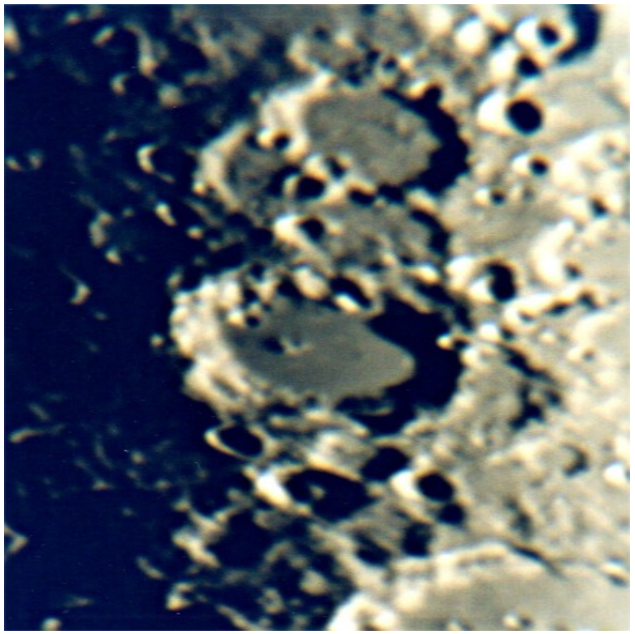


Campo aleatorio stazionario

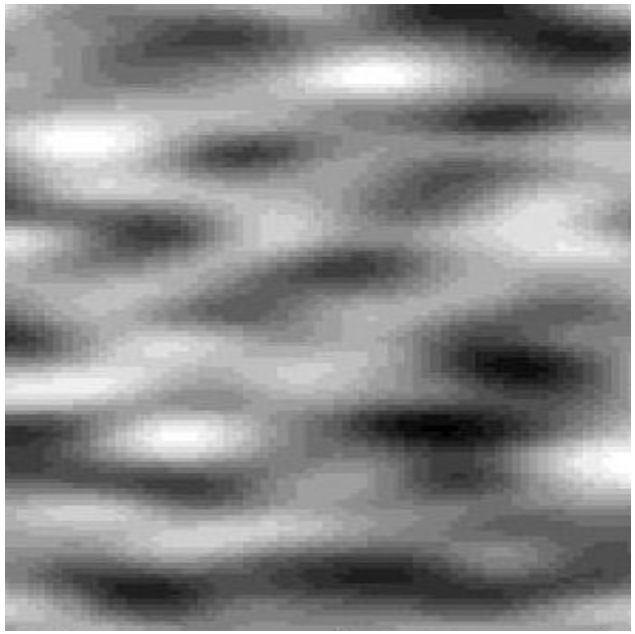
Modelli per la covarianza II

- Introduciamo un'ulteriore restrizione la covarianza dipende solo dalla distanza $\|h\| = \|s - s'\|$

$$C(\|s - s'\|) = \mathrm{Cov}(Z(s), Z(s')).$$



'Campi' lunari



Campo aleatorio isotropico

Variogramma (teorico)

- **Definizione**
  Si supponga che tutti i siti s_1 e s_2

$$\mathrm{E}(Z(s_1)) = \mathrm{E}(Z(s_2))$$

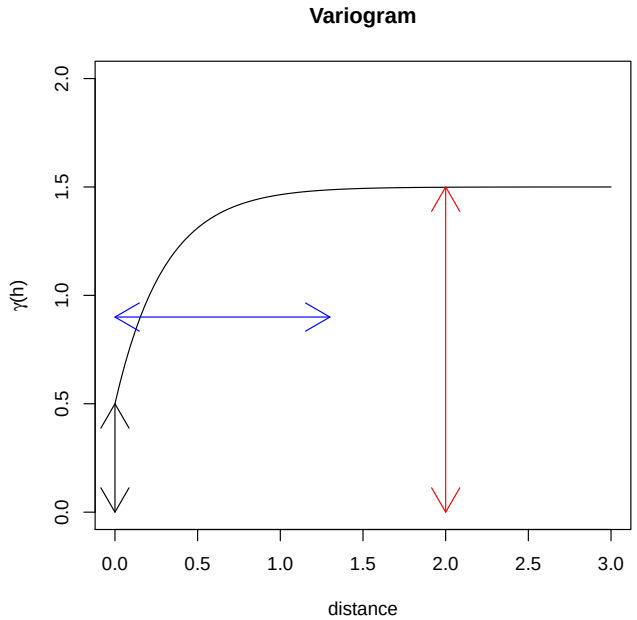
$$2\gamma(s_1 - s_2) = \mathrm{Var}(Z(s_1) - Z(s_2)),$$

  La quantità 2γ è detta variogramma e γ è detto semivariogramma.
- **Relazione tra variogramma e covariogramma**
  Se il processo è stazionario

$$\gamma(h) = (\sigma^2 - C(h)).$$

Interpretazione del (semi)variogramma

- *sill*: asintoto del semivariogramma. Solamente i processi stazionari in senso debole presentano un *sill*
- *range*: è la distanza h alla quale $\gamma(h)$ diviene costante ($C(h) \simeq 0$). Lo definiamo in maniera approssimata perché per alcuni processi la correlazione è zero solo asintoticamente.
- *nugget*: discontinuità del variogramma

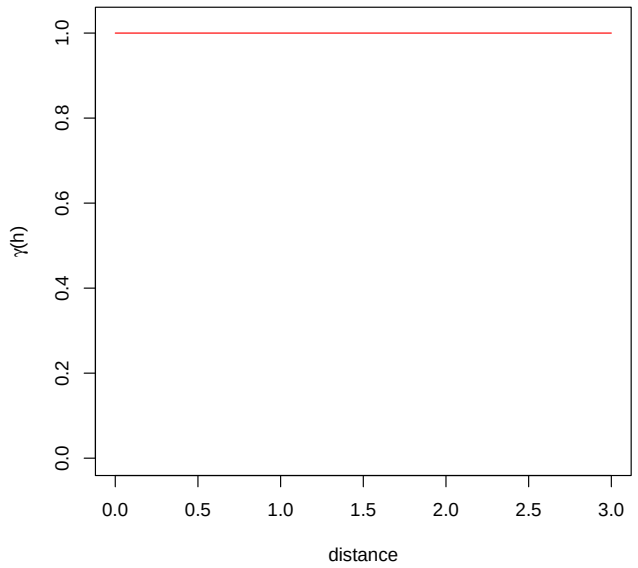


Esempi di (semi)variogramma (processi stazionari)

Modello pepita (*white noise*)

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \sigma^2 & h \neq 0 \end{cases}$$

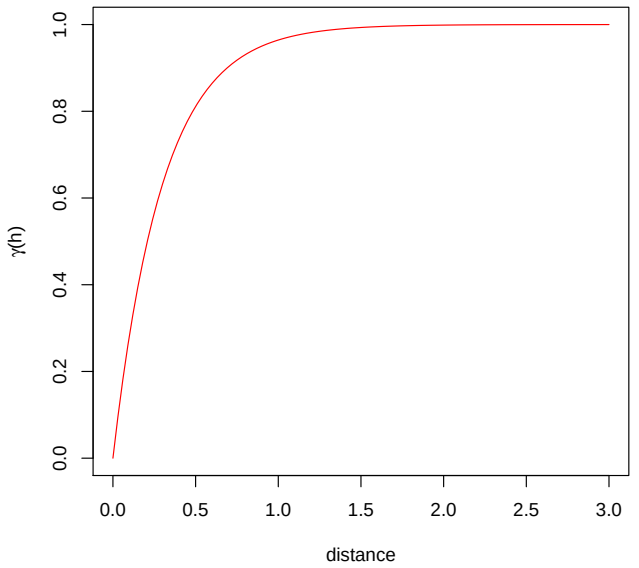
$$\sigma^2 = 1$$



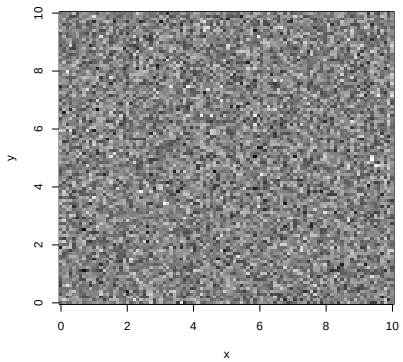
Modello esponenziale

$$\gamma(h) = \sigma^2\{1 - \exp\{-h/\phi\}\}$$

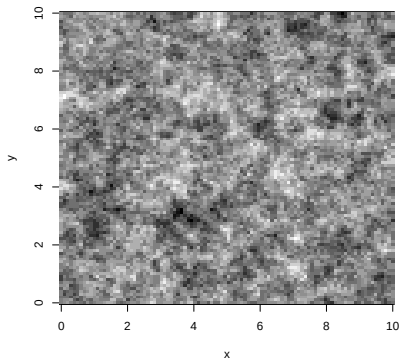
$$\sigma^2 = 1,\ \phi = 0.3$$



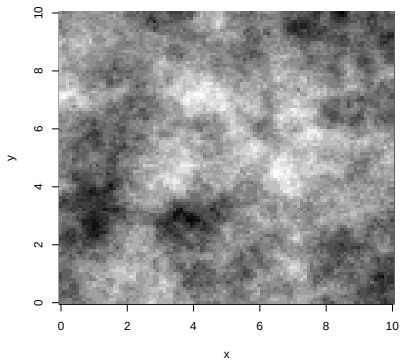
Regolarità



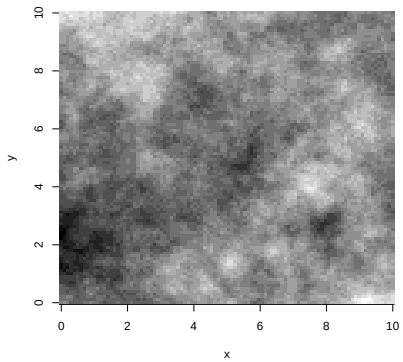
pepita



esponenziale ($\phi = 0.2$)



esponenziale ($\phi = 1$)



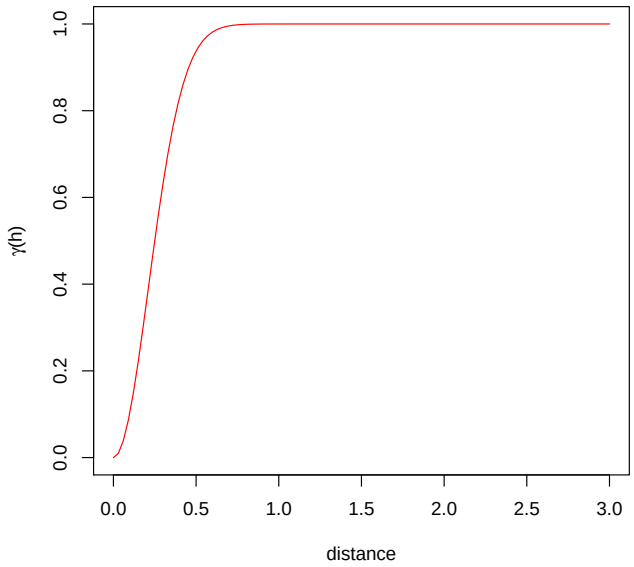
esponenziale ($\phi = 4$)

Esempi di (semi)variogramma (processi stazionari)

Modello gaussiano

$$\gamma(h) = \sigma^2\{1 - \exp\{-\frac{h^2}{\phi}\}\}$$

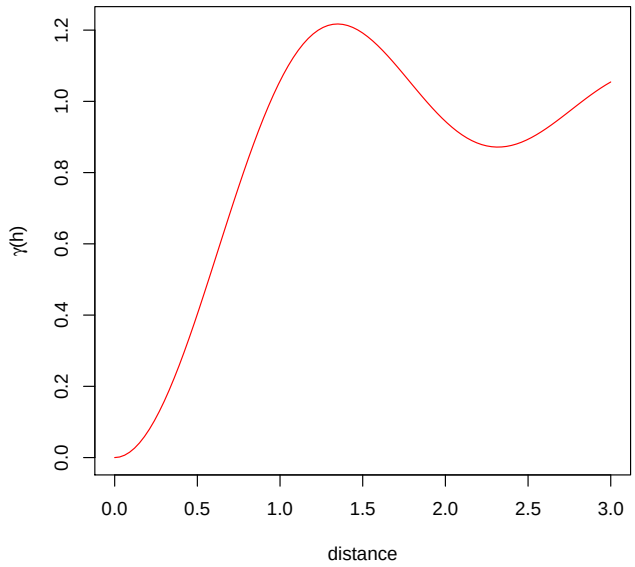
$$\sigma^2 = 1,\ \phi = 0.3$$



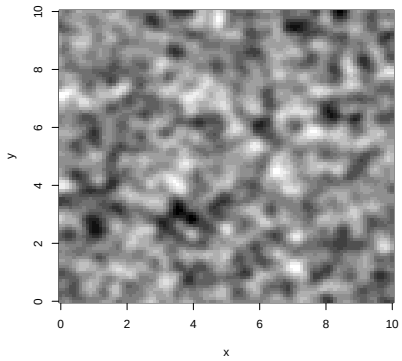
Modello *wave*

$$\gamma(h) = \sigma^2\{1 - \frac{\phi}{h}\sin(\frac{h}{\phi})\}$$

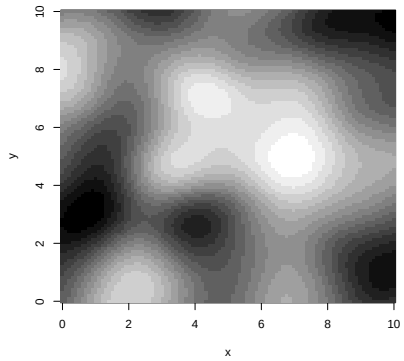
$$\sigma^2 = 1,\ \phi = 0.3$$



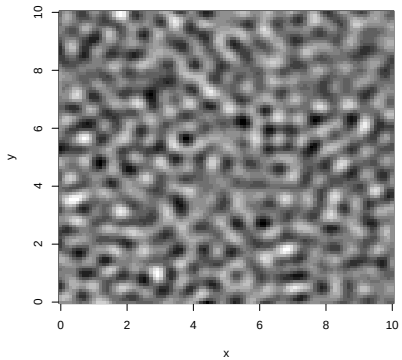
Regolarità



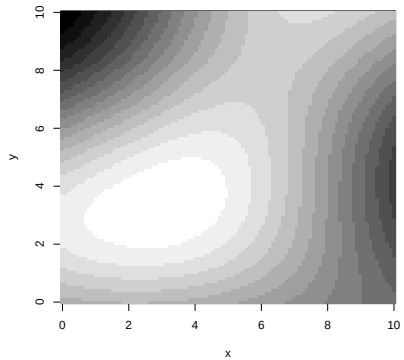
gaussiano ($\phi = 0.6/\sqrt{3}$)



gaussiano ($\phi = 2$)



wave ($\phi = .1$)



wave ($\phi = 2$)

Modelli di regressione: dati geostatistici

variabilità=larga scala+ piccola scala

$$Z(s) = \mu(s) + \varepsilon(s)$$

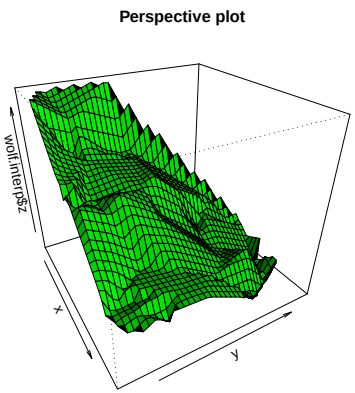
dove $\mu(s)$ componente di grande scala funzione **deterministica** sufficientemente 'liscia' e $\varepsilon(s)$ componente di piccola scala **aleatoria**.

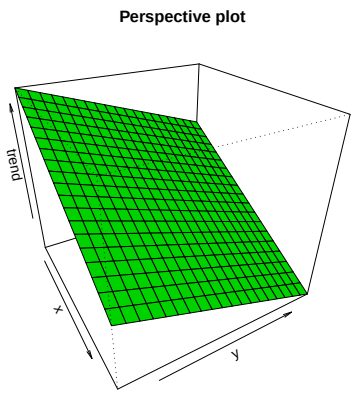
Superfici di trend e covariate

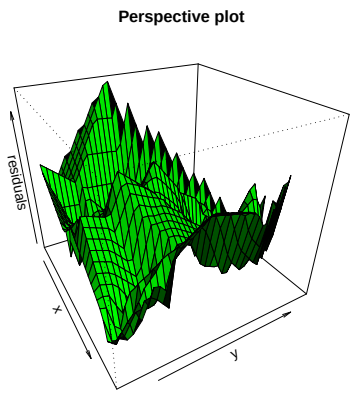
1. Superfici: la componente di larga scala dipende dalle coordinate
   - 1.1 $\mu(s) = \beta_0 + \beta_1 x + \beta_2 y, \qquad s = (x, y)'$
   - 1.2 $\mu(s) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$
2. Covariate

$$\mu(s) = \beta_0 + \beta_1 u(s).$$

Esempio: Acquifero del Texas



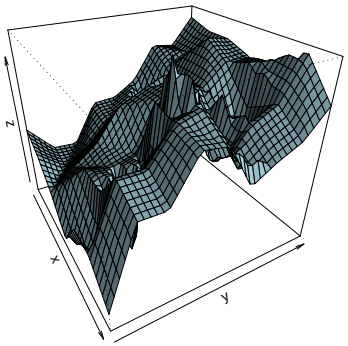


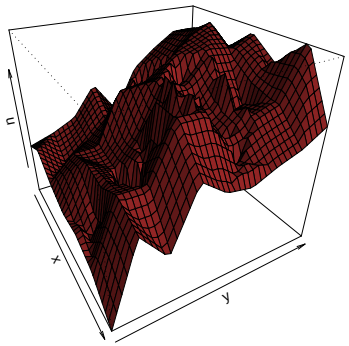


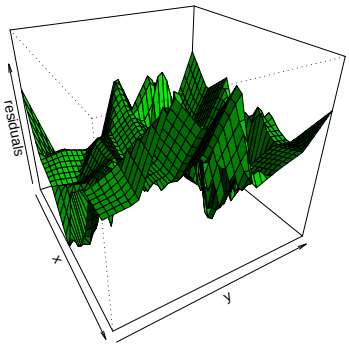
$$Z(s) = \mu(s) + \varepsilon(s)$$
$$Z(s) = \beta_0 + \beta_1 x + \beta_2 y + \varepsilon(s)$$

Esempio: Potenziali di ritenzione

Si considerano dei campioni prelevati dal suolo. Successivamente i campioni vengono saturati con acqua e si misura il potenziale di ritenzione ad una determinata pressione. Per ogni campione è nota la frazione granulometrica d'argilla.







$$Z(s) = \mu(s) + \varepsilon(s)$$
$$Z(s) = \beta_0 + \beta_1 u(s) + \varepsilon(s)$$

Previsione spaziale

- Problema: prevedere il valore $Z(s)$ quando s è un punto non campionato basandosi sui valori osservati $z(s_1), \ldots, z(s_n)$.
- La soluzione è detta anche interpolazione
- Obbiettivo: costruzione di mappe

Si suppone che il modello sia dato da

$$Z(s) = \mu(s) + \varepsilon(s).$$

Due approcci

1. Proporre un modello per $\mu(s)$, stimarlo ($\widehat{m}(s)$), sottrarlo dai dati $\widehat{\varepsilon}(s) = Z(s) - \widehat{m}(s)$, quindi modellare e prevedere i residui $\widehat{\varepsilon}(s)$
2. Modellare e stimare simultaneamente $\mu(s)$ e $\varepsilon(s)$

Kriging

- Previsore lineare $\widehat{Z}(s)$ non distorto ottimo secondo il criterio dell'errore quadratico medio

$$\mathrm{E}[Z(s) - \widehat{Z}(s)]^2$$

- E' un previsore ottimo solo rispetto al criterio e al modello scelto
- Basato sulla teoria dei processi stocastici con funzioni di covarianza

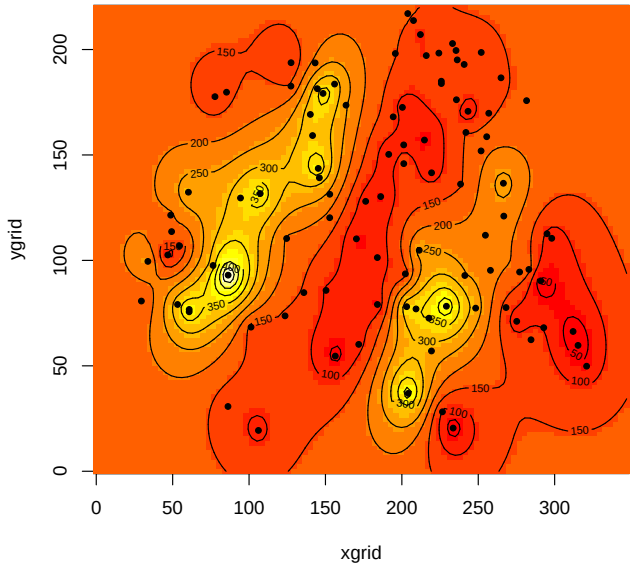
Kriging: caratteristiche

- Il previsore in ogni punto è una media pesata dei valori osservati $z(s_1), \ldots, z(s_n)$

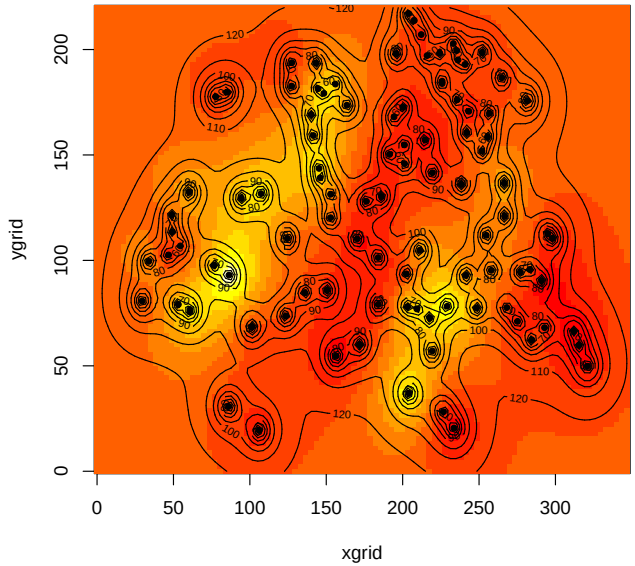
$$\widehat{Z}(s) = \lambda_1 z(s_1) + \cdots + \lambda_n z(s_n)$$

- L'errore quadratico medio di previsione è calcolato automaticamente nella procedura di calcolo.
- La superficie prevista è liscia il grado di lisciamento dipende dal variogramma.
- I punti campionati sono previsti esattamente.

Kriging: esempio precipitazione in Svizzera

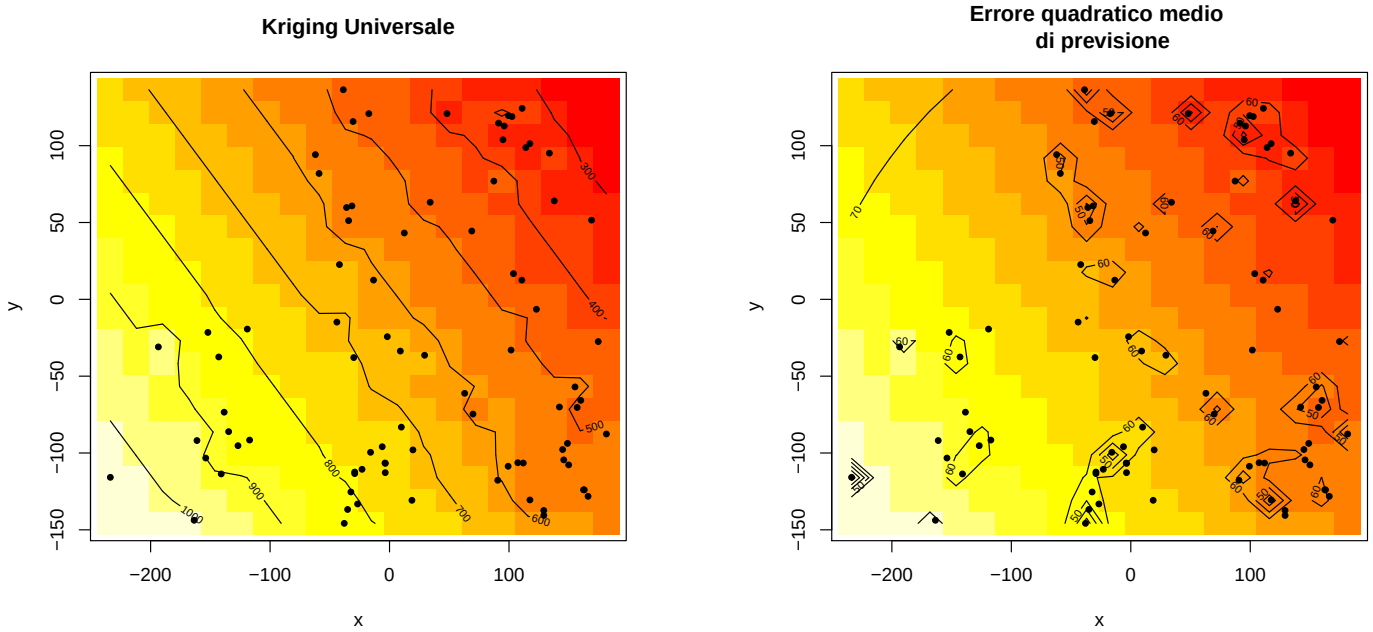


Kriging



$\sqrt{\text{Errore quadratico medio}}$

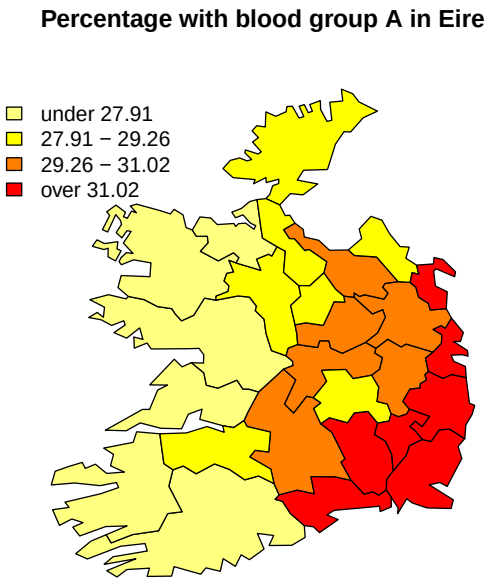
Kriging: esempio acquifero del Texas

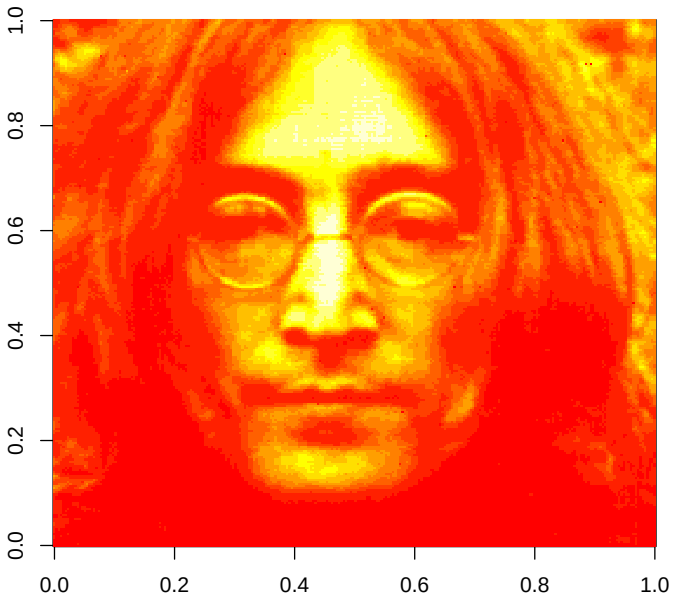


Dati di area

Dati di area: notazioni

Una regione D è suddivisa in n aree A_i, $i = 1, \ldots, n$ e viene osservato il valore su ogni area $Z_i = Z(A_i)$, $i \in S = \{1, 2, \cdots, n\}$.

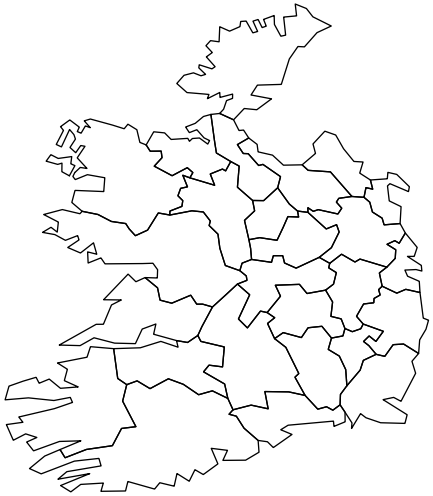




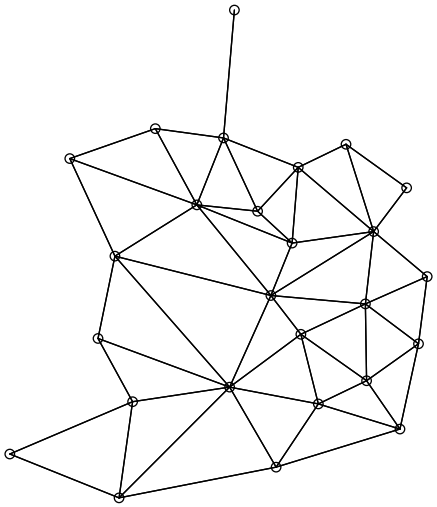
Misure di vicinanza I

- Distanze tra i centroidi (dati continui)
- Matrice di prossimità ($\mathbf{W} = (w_{ij})$ di dimensione $n \times n$. Esempi:

Percentage with blood group A in Eire



Graph



Misure di vicinanza II

- Matrice di contiguità (induce un grafo)

$$w_{ij} = \begin{cases} 1 & \text{se } A_i \text{ ha un confine in comune con } A_j \\ 0 & \text{altrimenti} \end{cases}$$

- W può essere non simmetrica

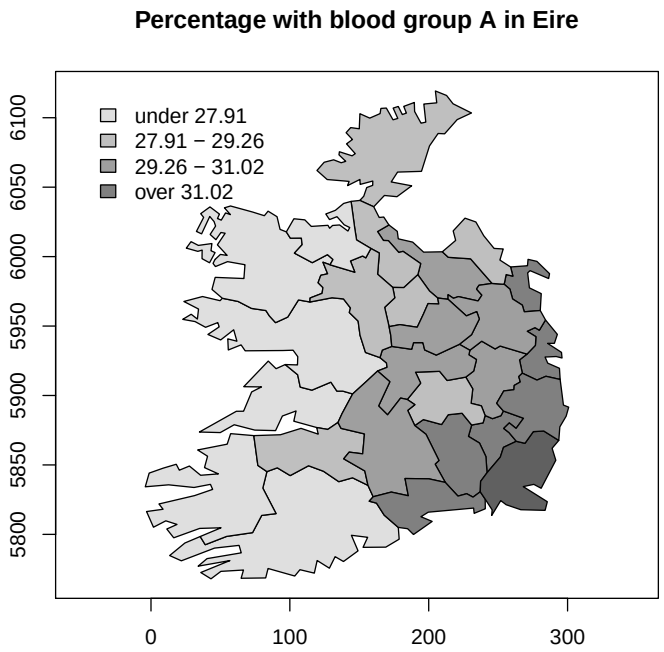
$$w_{ij} = \frac{l_{ij}}{l_i}$$

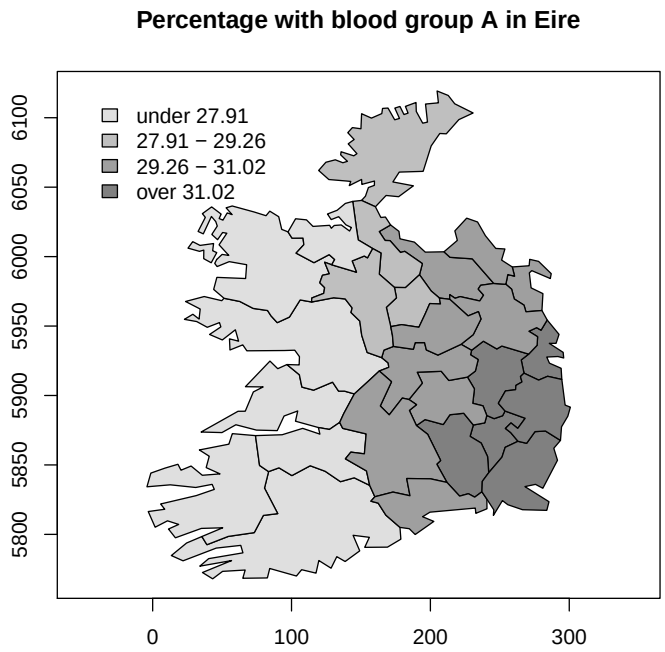
  dove l_{ij} è la lunghezza della frontiera in comune tra A_i e A_j e $l_i = \sum_j l_{ij}$ è il perimetro di A_i
- $\mathbf{W}^{(k)}$ matrice di distanza al ritardo spaziale k ad esempio $w_{ij}^{(2)} = 1$ se A_i e A_j non sono confinanti ma confinano tutte e due con una regione A_k.

Lisciamento

Medie mobili spaziali

$$\hat{\mu}_i = \frac{z_i + \sum_{j=1}^{n} w_{ij} z_j}{1 + \sum_{j=1}^{n} w_{ij}}$$





Autocorrelazione e sue misure

Una W-misura dell' autocovarianza spaziale è data da

$$C = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}}$$

Questa autocovarianza deve essere normalizzata per ottenere una autocorrelazione. Ad esempio con la varianza campionaria aggiustata

$$s_*^2 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})^2}{n} = \frac{(n-1)s^2}{n}$$

Ciò ci conduce alla definizione dell'indice di Moran

Indice di Moran

$$M = \frac{C}{s_*^2} = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(z_i - \overline{z})(z_j - \overline{z})}{\{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}\} \sum_{i=1}^{n} (z_i - \overline{z})^2}$$

- L'indice di Moran M può uscire dall'intervallo $[-1, +1]$ (non è proprio un'autocorrelazione !)
- È tanto più grande quanto valori in siti vicini sono simili (è tanto più piccolo quanto valori in siti vicini non sono simili) $\rightarrow$ aggregazione o autocorrelazione positiva (repulsione o autocorrelazione negativa)
- Il caso in cui le osservazioni sono indipendenti corrisponde ad un valore di I_n^M vicino a 0.

Nel nostro caso $M = 0.554$.

Indice di Geary

Riprendiamo la misura di dissimilarità $(z_i - z_j)^2$.
Una misura media

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(z_i - z_j)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Una misura media standardizzata è l'*indice di Geary*

$$G = \frac{D}{s^2} = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(z_i - z_j)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ \sum_{i=1}(z_i - \overline{z})^2}$$

- assume valori nell'intervallo $[0, 2]$.
- valori prossimi a 1 indicano assenza di autocorrelazione. Valori inferiori (superiori) a 1 indicano autocorrelazione positiva (negativa)
- G assume valori vicino allo zero quando a siti vicini corrispondono valori simili (legge di Tobler)

Nel nostro caso $G = 0.38$